Is Space-Time Attention All You Need for Video Understanding?

Supplementary Materials

Gedas Bertasius¹  Heng Wang¹  Lorenzo Torresani¹²

Our supplementary materials consist of:

1. Implementation Details.
2. Additional Ablations.
3. Additional Qualitative Results.

1. Implementation Details

Our TimeSformer implementation is built using PySlowFast (Fan et al., 2020) and pytorch-image-models (Wightman, 2019) packages. Below, we describe specific implementation details regarding the training and inference procedures of our model.

**Training.** We train our model for 15 epochs with an initial learning rate of 0.005, which is divided by 10 at epochs 11, and 14. During training, we first resize the shorter side of the video to a random value in [256, 320]. We then randomly sample a 224 × 224 crop from the resized video. For our high-resolution model, TimeSformer-HR, we resize the shorter side of the video to a random value in [448, 512], and then randomly sample a 448 × 448 crop. We randomly sample clips from the full-length videos with a frame rate of 1/32. The batch size is set to 16. We train all our models using synchronized SGD across 32 GPUs. The momentum is set to 0.9. A dropout of 0.5 is used before the final classification layer. The final prediction is obtained by averaging the softmax scores of these 3 predictions.

**Datasets.** Kinetics-400 (Carreira & Zisserman, 2017) consists of 240K training videos and 20K validation videos that span 400 human action categories. Kinetics-600 (Carreira et al., 2018) has 392K training videos and 30K validation videos spanning 600 action categories. Something-Something-V2 (Goyal et al., 2017b) contains 170K training videos and 25K validation videos that span 174 action categories. Lastly, Diving-48 (Li et al., 2018) has 16K training videos and 3K testing videos spanning 48 fine-grained diving categories. For all of these datasets, we use standard classification accuracy as our main performance metric.

2. Additional Ablations

**Smaller & Larger Transformers.** In addition to the “Base” ViT model (Dosovitskiy et al., 2020), we also experimented with the “Large” ViT. We report that this yielded results 1%
worse on both Kinetics-400, and Something-Something-V2. Given that our “Base” model already has 121M parameters, we suspect that the current datasets are not big enough to justify a further increase in model capacity. We also tried the “Small” ViT variant, which produced accuracies about 5% worse than our default “Base” ViT model.

**Larger Patch Size.** We also experimented with a different patch size, i.e., $P = 32$. We report that this variant of our model produced results about 3% worse than our default variant using $P = 16$. We conjecture that the performance decrease with $P = 32$ is due to the reduced spatial granularity. We did not train any models with $P$ values lower than 16 as those models have a much higher computational cost.

**The Order of Space and Time Self-Attention.** Our proposed “Divided Space-Time Attention” scheme applies temporal attention and spatial attention one after the other. Here, we investigate whether reversing the order of time-space attention (i.e., applying spatial attention first, then temporal) has an impact on our results. We report that applying spatial attention first, followed by temporal attention leads to a 0.5% drop in accuracy on both Kinetics-400, and Something-Something-V2. We also tried a parallel space-time self-attention. We report that it produces 0.4% lower accuracy compared to our adopted “Divided Space-Time Attention” scheme.

### 3. Additional Qualitative Results

In Figure 1, we present space-time attention visualizations obtained by applying TimeSformer on Something-Something-V2 videos. To visualize the learned attention, we use the Attention Rollout scheme presented in (Abnar & Zuidema, 2020). Our results suggest that TimeSformer learns to attend to the relevant regions in the video in order to perform complex spatiotemporal reasoning. For example, we can observe that the model focuses on the configuration of the hand when visible and the object-only when not visible.

### References


Figure 1. Visualization of space-time attention from the output token to the input space on Something-Something-V2. Our model learns to focus on the relevant parts in the video in order to perform spatiotemporal reasoning.